计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300271

基于深度神经网络和自注意力机制的医学实体关系抽取

张世豪　杜圣东　贾　真　李天瑞
西南交通大学计算机与人工智能学院　成都 611756
(shihao_zura@163.com)

摘　要　随着医学信息化的推进,医学领域已经积累了海量的非结构化文本数据,如何从这些医学文本中挖掘出有价值的信息,是医学行业和自然语言处理领域的研究热点。随着深度学习的发展,深度神经网络被逐步应用到关系抽取任务中,其中"recurrent+CNN"网络框架成为了医学实体关系抽取任务中的主流模型。但由于医学文本存在实体分布密度较高、实体之间的关系交错互联等问题,使得"recurrent+CNN"网络框架无法深入挖掘医学文本语句的语义特征。基于此,在"recurrent+CNN"网络框架基础之上,提出一种融合多通道自注意力机制的中文医学实体关系抽取模型,包括:1)利用BLSTM捕获文本句子的上下文信息;2)利用多通道自注意力机制深入挖掘句子的全局语义特征;3)利用CNN捕获句子的局部短语特征。通过在中文医学文本数据集上进行实验,验证了该模型的有效性,其精确率、召回率和F1值与主流的模型相比均有提高。

关键词:医学文本;实体关系抽取;多通道自注意力;深度学习
中图法分类号　TP391

Medical Entity Relation Extraction Based on Deep Neural Network and Self-attention Mechanism

ZHANG Shi-hao,DU Sheng-dong,JIA Zhen and LI Tian-rui
School of Computing and Artificial Intelligence,Southwest Jiaotong University,Chengdu 611756,China

Abstract　With the advancement of medical informatization,a large amount of unstructured text data has been accumulated in the medical field. How to mine valuable information from these medical texts is a research hotspot in the field of medical profession and natural language processing. With the development of deep learning,deep neural network is gradually applied to relation extraction task,and "recurrent+CNN" network framework has become the mainstream model in medical entity relation extraction task. However,due to the problems of high entity density and the cross-connection of relationships between entities in medical texts,the "recurrent+CNN" network framework cannot deeply mine the semantic features of medical texts. Based on the "recurrent+CNN" network framework,this paper proposes a Chinese medical entity relation extraction model with multi-channel self-attention mechanism. It includes that BLSTM is used to capture the context information of text sentences,a multi-channel self-attention mechanism is used to mine the global semantic features of sentences,and CNN is used to capture the local phrase features of sentences. The effectiveness of the model is verified by experiments on Chinese medical text dataset. The precision,recall and F1 value of the model are improved compared with the mainstream models.

Keywords　Medical text,Entity relation extraction,Multi-channel self-attention,Deep learning

1　引言

大数据时代下,信息技术的快速发展与广泛应用推动医学领域朝着医学信息化的方向发展,并成为了一种主流趋势。随着医学信息化的推进,医学领域已经积累了海量的非结构化文本数据,其中包含了大量有价值的信息,如疾病、症状、药物、检查等重要的医学实体,以及各医学实体之间丰富的语义关系。如何从这些医学文本中挖掘出有效的信息并加以存储管理,以构建大规模、高质量的医学知识图谱,对医学信息化的发展具有重大意义,也是自然语言处理(Natural Language Processing,NLP)领域的研究热点。

实体关系抽取(entity relation extraction)作为信息抽取[1]的核心任务之一,旨在从非结构化文本中自动地抽取出实体对之间的语义关系,从而提取出有效的语义信息,在知识图谱的构建中有着十分重要的地位[2]。该任务最早由信息理解会议(Message Understanding Conference,MUC)[3]引入,

到稿日期:2021-03-29　返修日期:2021-05-21　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:四川省重点研发项目(2020YFG0035)
This work was supported by the Sichuan Key R&D Project(2020YFG0035).
通信作者:李天瑞(trli@swjtu.edu.cn)

使关系抽取在通用领域中取得了许多研究成果。直到2010年美国国家集成生物与临床信息学研究中心(Informatics for Integrating Biology and the Bedside, I2B2)发布了基于英文电子病历的医学实体关系抽取任务[4],才使得医学实体关系抽取成为了研究热点。不过在中文医学实体关系抽取方面,公开的评测以及相应的研究还相对较少。

传统的医学实体关系抽取方法包括基于规则的方法、基于特征向量的方法和基于核函数的方法。这些方法虽然都取得了一定的效果,但需要依赖人工设计的规则或特征,从而使得模型的性能取决于人工手动设计的规则或特征的质量。近年来,随着深度学习的发展,以神经网络为主的方法被应用到关系抽取任务中,并取得了很多成果。该方法不依赖人工设计特征,完全由神经网络自动学习相关的所有特征。目前,以卷积神经网络(Convolution Neural Network, CNN)和循环神经网络(Recurrent Neural Network, RNN)为代表的深度学习方法在医学实体关系抽取任务上取得了突破。CNN能够捕获语句中的局部信息,但忽略了语句中全局信息的作用;RNN可以有效学习文本语句的上下文依赖信息,但无法挖掘出句法和语义层面的特征。以RNN和CNN相结合的"recurrent+CNN"网络框架是医学实体关系抽取任务中的主流模型[5],可以同时捕获语句的上下文信息和局部信息。由于医学领域的特殊性,医学文本存在着实体分布密度较高、实体之间的关系交错互联等现象,使得"recurrent+CNN"网络框架无法深入挖掘医学文本语句的语义特征。例如,对于医学文本语句"颈静脉怒张、p2 亢进、下肢水肿和肝肿大是肺心病的体征,持续氧疗是主要治疗",存在着3种类型共6个实体,实体之间有5组关系,分别为(肺心病,临床表现,颈静脉怒张),(肺心病,临床表现,p2 亢进),(肺心病,临床表现,下肢水肿),(肺心病,临床表现,肝肿大),(肺心病,辅助治疗,持续氧疗)。医学文本中普遍存在这种实体分布密度较高,且实体之间的关系也交错互联的现象。

基于此,本文提出一种融合多通道自注意力机制的中文医学实体关系抽取模型BLSTM-MCatt-CNN。该模型采用"recurrent+CNN"网络框架,其中的recurrent利用双向长短期记忆网络(Bidirectional Long Short-Term Memory Network, BLSTM)来捕获医学文本语句的上下文信息和浅层语义特征,利用CNN捕获医学文本语句的局部短语特征,并结合多通道自注意力机制(Multi-Channel Self-Attention, MCatt)捕获医学文本语句的全局信息,从而对医学文本的语义特征进行深入挖掘。

2 相关工作

关系抽取方法在早期是利用基于规则[6]、基于特征向量和基于核函数[7]的方法,且大多数是在通用领域中进行的。基于规则的方法使用句子分析工具来识别文本语句中的句法元素,然后根据这些元素构建模式规则,并根据规则进行关系抽取。基于特征向量的方法主要是根据特征向量的相似度训练支持向量机[8]、最大熵[9]、条件随机场[10]等机器学习模型来进行关系抽取。基于核函数的方法通过设计特定核函数来计算句子之间的相似度,再根据相似度对关系进行分类。而在医学领域,Rink等[11]使用单个支持向量机分类器来识别医学实体之间的关系;D'Souza等[12]采用特征向量和核函数集成的方法进行实体关系抽取;Kim等[13]使用基于特征的线性核函数方法抽取药物之间的关系。

随着近年来深度学习的发展,许多研究者将神经网络应用于关系抽取任务中,并在通用领域取得了突破。其中常用的神经网络模型包括CNN[14]、RNN[15]及其变种LSTM[16]。考虑到句子中每个单词对关系语义的贡献程度不同,研究者们引入了注意力机制,将其分别与CNN[17]和LSTM[18-19]进行结合,取得了不错的效果。RNN模型可以有效学习文本序列的上下文依赖关系,但无法挖掘出句法和语义层面的特征。CNN模型能够捕获语句中的局部信息,但忽略了全局信息的作用。基于此,Cai等[20]将上述两种模型相结合,提出了BRCNN(Bidirectional Recurrent CNN)模型,以同时捕获语句的上下文信息和局部信息。Zhang等[21]和Tran等[22]将注意力机制引入BLSTM和CNN的混合模型中,从而对句子级别的特征进行进一步的学习。

深度学习也在医学领域的关系抽取任务中得到了应用。Sahu等[23]采用CNN对临床医学文本进行关系抽取,证明了深度学习的方法在医学领域同样适用。Zhou等[24]将先验医学知识引入CNN中对化学疾病关系进行自动地提取。Sahu等[25]提出了一种基于BLSTM的药物相互作用提取模型,并分别采用最大池化与注意力池化对文本的上下文特征进行提取。Bai等[26]提出了一种基于多层注意力的LSTM关系抽取模型,用于从医学短文本中提取非特定的关系。后来,许多研究者采用"recurrent+CNN"网络框架,并通过实验验证了其在医学实体关系抽取任务上的有效性。如Raj等[27]与He等[28]分别将BLSTM、双向门控循环单元(Bidirectional Gated Recurrent Unit, BGRU)网络和CNN混合进行关系抽取,实验结果表明,这种方法比单纯使用CNN或RNN的方法效果更好。

3 模型介绍

针对医学文本存在的实体分布密度较高、实体之间的关系交错互联导致的抽取效果不佳的问题,本文提出了一种融合多通道自注意力机制的中文医学实体关系抽取模型BLSTM-MCatt-CNN,其结构如图1所示,主要分为以下5个部分。

(1)嵌入层:对输入的医学文本语句采用向量表示技术进行向量化,将语句中的每个字转换为由字向量和位置向量拼接而成的低维稠密实值向量,得到语句的输入特征向量。

(2)BLSTM层:利用BLSTM从输入特征向量中学习文

本语句的上下文信息和浅层语义特征，得到句子向量。

(3)多通道自注意力层：利用多通道自注意力机制，从句子向量中学习文本语句的深层次全局语义特征，得到句子的全局特征向量。

(4)CNN 层：利用 CNN 从句子向量中学习文本语句的局部短语特征，得到句子的局部特征向量。

(5)输出层：将句子的全局特征向量和局部特征向量进行拼接，输入到一个全连接网络中，最后经过 softmax 函数输出结果。

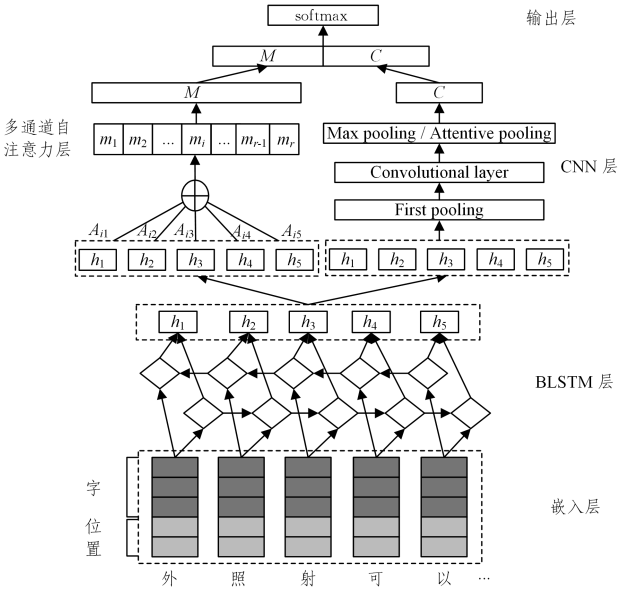

图 1 BLSTM-MCatt-CNN 模型

Fig. 1 BLSTM-MCatt-CNN model

接下来对模型的 5 个部分分别进行详细的阐述。

3.1 嵌入层

对于一个含有两个标记实体的医学文本语句，使用字符特征和字相对于实体的位置特征对该语句的每个字进行信息表征，并利用向量化技术将每个字的字符特征和位置特征映射为低维稠密实值向量，从而得到整个语句的输入特征向量。

字向量：对于语句序列 $X=(x_1,x_2,\cdots,x_n)$，将序列中的每个字 x_i 通过字向量矩阵 W_{char} 转换成一个低维的稠密实值向量 $\boldsymbol{w}_i^{\text{char}}$：

$$\boldsymbol{w}_i^{\text{char}}=\boldsymbol{W}_{\text{char}}\boldsymbol{v}_i \tag{1}$$

其中，字向量矩阵 $\boldsymbol{W}_{\text{char}}$ 的维度为 $|V|\times d^{\text{char}}$，$|V|$ 是输入字表的大小，d^{char} 是字向量的维度，$\boldsymbol{v}_i$ 是字 x_i 在字表中的 one-hot 编码。本文使用 Word2vec 工具训练字向量。

位置向量：使用位置特征表示每个字相对于实体所处的位置。对于语句序列 X 中的每个字 x_i，计算其到实体的距离，计算方式如下：

$$p_i^j=p_i-e_j^p,i\in[1,2,\cdots,n],j\in[1,2] \tag{2}$$

其中，p_i 表示字 x_i 在语句序列 X 中的位置，e_j^p 表示实体 e_j 在语句序列 X 中的位置。将每个位置特征映射到位置特征空间中一个随机初始化的位置向量上，维度为 d^p。每个字 x_i 都包含两个位置向量 $\boldsymbol{w}_i^{p,1}$，$\boldsymbol{w}_i^{p,2}$。

最后，将每个字的字向量 $\boldsymbol{w}_i^{\text{char}}$ 与两个位置向量 $\boldsymbol{w}_i^{p,1}$ 和 $\boldsymbol{w}_i^{p,2}$ 进行拼接，得到每个字 x_i 最终的输入特征向量 $\boldsymbol{w}_i=[\boldsymbol{w}_i^{\text{char}},\boldsymbol{w}_i^{p,1},\boldsymbol{w}_i^{p,2}]$，维度为 $d=d^{\text{char}}+2\,d^p$。于是，整个语句序列的输入特征向量可以表示为 $\boldsymbol{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n)$。

3.2 BLSTM 层

文本数据可以被视为具有前后依赖关系的序列数据。LSTM 作为 RNN 的一个变种，适合处理序列数据。LSTM 通过设置输入门 i、遗忘门 f 和输出门 o 这 3 个门控机制来控制信息流，并结合细胞状态 c 实现对历史信息的更新、取舍和存储，解决了 RNN 中存在的梯度消失问题。一个标准的 LSTM 在 t 时刻的隐藏层状态输出 h_t 的计算方式如下：

$$i_t=\sigma(\boldsymbol{W}_{wi}\boldsymbol{w}_t+\boldsymbol{W}_{hi}h_{t-1}+\boldsymbol{W}_{ci}\circ c_{t-1}+\boldsymbol{b}_i) \tag{3}$$

$$f_t=\sigma(\boldsymbol{W}_{wf}\boldsymbol{w}_t+\boldsymbol{W}_{hf}h_{t-1}+\boldsymbol{W}_{cf}\circ c_{t-1}+\boldsymbol{b}_f) \tag{4}$$

$$c_t=f_t\circ c_{t-1}+i_t\circ\tanh(\boldsymbol{W}_{wc}\boldsymbol{w}_t+\boldsymbol{W}_{hc}h_{t-1}+\boldsymbol{b}_c) \tag{5}$$

$$o_t=\sigma(\boldsymbol{W}_{wo}\boldsymbol{w}_t+\boldsymbol{W}_{ho}h_{t-1}+\boldsymbol{W}_{co}\circ c_t+\boldsymbol{b}_o) \tag{6}$$

$$h_t=o_t\circ\tanh(c_t) \tag{7}$$

其中，i_t，f_t，o_t 分别表示 t 时刻的输入门、遗忘门和输出门，$\boldsymbol{w}_t$ 表示当前时刻 t 的输入数据，h_{t-1} 表示前一时刻 LSTM 隐藏层状态，c_{t-1} 和 c_t 分别表示前一时刻和当前时刻的细胞状态，$\sigma(\cdot)$ 代表 sigmoid 函数，$\circ$ 表示 Hadamard 乘积，$\boldsymbol{W}_*$ 为对应门控机制中需要学习的权重矩阵，$\boldsymbol{b}_*$ 为对应门控机制中的偏置向量。

然而，标准的 LSTM 只能获取序列数据在某一时刻的上文信息，却无法获取该时刻的下文信息。而 BLSTM 通过使用一个前向 LSTM 和一个后向 LSTM 来分别捕获序列数据的上文信息和下文信息，通过将两个 LSTM 的隐藏层输出进行拼接，从而使模型具备捕获文本数据上下文信息的能力。在 t 时刻 BLSTM 的隐藏层状态输出如下：

$$h_t=[\overrightarrow{h_t},\overleftarrow{h_t}] \tag{8}$$

其中，$\overrightarrow{h_t}$ 和 $\overleftarrow{h_t}$ 分别表示前向 LSTM 和后向 LSTM 在 t 时刻的隐藏层状态输出。对嵌入层得到的输入特征向量 $\boldsymbol{W}$ 进行 BLSTM 编码后，可以得到一个句子向量 $\boldsymbol{H}=(h_1,h_2,\cdots,h_n)$，维度为 $n\times 2u$，其中 u 表示单层 LSTM 的隐藏层神经元个数。

3.3 多通道自注意力层

传统的自注意力机制将句子中的每个字和该句子中的所有字进行注意力权重计算，得到一组注意力权重向量，从而学习到句子内部字之间的长距离依赖关系，并捕获句子的全局结构特征。然而，这种自注意力机制所得到的权重向量通常只能表示句子的某一个方面，而医学文本具有实体分布密度较高的特点，因此对于一个医学文本语句，可能存在多个方面共同构成句子的整体语义。若对句子进行多次注意力权重计算，则能得到多组不同的注意力权重向量，从而可以从多个方面完整地表示语句，这就是多通道自注意力机制[29]。多通道自注意力机制将传统自注意力的权重向量改为权重矩阵，其

中权重矩阵的一行就代表一个方面。因此，本文采用多通道自注意力机制来捕获医学文本语句的多语义全局信息，对医学文本的语义特征进行深入挖掘。

具体来讲，将 BLSTM 层得到的句子向量 $\boldsymbol{H}=(h_1,h_2,\cdots,h_n)$ 作为自注意力机制的输入。传统的自注意力机制权重的计算公式如下：

$$\boldsymbol{a}=\text{softmax}(\boldsymbol{w}_{s2}\tanh(\boldsymbol{W}_{s1}\boldsymbol{H}^{\mathrm{T}})) \tag{9}$$

其中，$\boldsymbol{W}_{s1}$ 是维度为 $d_a\times 2u$ 的权重矩阵，$\boldsymbol{w}_{s2}$ 是维度为 d_a 的权重向量，$\boldsymbol{a}$ 是维度为 n 的注意力权重向量。多通道自注意力机制将权重向量 $\boldsymbol{w}_{s2}$ 扩展成维度为 $r\times d_a$ 的权重矩阵 $\boldsymbol{W}_{s2}$，其中 r 表示通道数，用于从不同方面获取句子的语义信息。多通道自注意力机制权重的计算公式如下：

$$\boldsymbol{A}=\text{softmax}(\boldsymbol{W}_{s2}\tanh(\boldsymbol{W}_{s1}\boldsymbol{H}^{\mathrm{T}})) \tag{10}$$

其中，注意力权重向量 $\boldsymbol{a}$ 也变成了维度为 $r\times n$ 的权重矩阵 $\boldsymbol{A}$。句子向量 $\boldsymbol{H}$ 根据注意力权重 $\boldsymbol{A}$ 做加权求和，得到语句的全局特征向量：

$$\boldsymbol{M}=\boldsymbol{A}\boldsymbol{H} \tag{11}$$

其中，$\boldsymbol{M}$ 为得到的全局特征向量，维度为 $r\times 2u$。

3.4 CNN 层

BLSTM 层输出的句子向量 $\boldsymbol{H}$ 包含了句子的上下文信息。然而在大多数情况下，一个句子中的某些字符对整个句子的表达可能并不重要。因此，本文在 BLSTM 层后使用最大池化技术从句子的多个短语中提取最重要的特征。将 BLSTM 层输出的句子向量 $\boldsymbol{H}=(h_1,h_2,\cdots,h_n)$ 作为该池化层的输入，最大池化的计算公式如下：

$$p_i=\max\{h_{i+1},h_{i+2},\cdots,h_{i+f_1}\} \tag{12}$$

$$p=(p_1,p_2,\cdots,p_{n-f_1+1}) \tag{13}$$

其中，f_1 表示用于池化的滤波器的长度，n 表示输入句子的长度，p_i 表示句子向量中第 i 个长度为 f_1 的短语中的最大值，p 表示整个句子最大池化后的输出。接着，对池化层的输出 p 进行卷积操作，从而捕获句子中每个短语部分的局部特征。卷积层的计算公式如下：

$$h_c^i=f(\boldsymbol{w}_c\cdot p^{i:i+f_2-1}+\boldsymbol{b}_c) \tag{14}$$

其中，$\boldsymbol{w}_c$ 表示权重向量，$\boldsymbol{b}_c$ 表示偏置项，f 表示 ReLU 函数，f_2 表示卷积层滤波器的长度。如果卷积层滤波器的个数为 n_c，则通过卷积层可以得到一个维度为 $n_c\times(n-f_1-f_2+2)$ 的输出矩阵 $\boldsymbol{H}_c$。卷积层的输出长度 $(n-f_1-f_2+2)$ 取决于输入语句的长度 n，具有可变性。为了获得整个句子的固定长度的特征，我们在卷积层后进行第二次池化。这里分别采用最大池化和注意力池化两种不同的池化模式。

3.4.1 最大池化

最大池化通过获取整个句子中的最大值，从而捕获最重要的特征。由于该层的输入是局部卷积向量，因此该层本质上是从句子的几个短语中提取最重要的局部特征。最大池化的计算公式如下：

$$h_{\text{pool}}=\max\{h_c^1,h_c^2,\cdots,h_c^{n-f_1-f_2+2}\} \tag{15}$$

3.4.2 注意力池化

如果重要特征分布在句子的不同短语中，则最大池化可能导致一些信息丢失。因此，采用一个基于注意力的池化模式来解决这个问题，该模式通过向量的加权线性组合来获得最佳的特征向量[18]。将卷积层的输出矩阵 $\boldsymbol{H}_c$ 作为输入。注意力池化的计算公式如下：

$$\boldsymbol{\alpha}=\text{softmax}(\boldsymbol{w}_a\tanh(\boldsymbol{H}_c)) \tag{16}$$

$$h_{\text{att}}=\boldsymbol{\alpha}\boldsymbol{H}_c^{\mathrm{T}} \tag{17}$$

其中，$\boldsymbol{w}_a$ 是维度为 n_c 的权重向量；$\boldsymbol{\alpha}$ 是注意力权重向量，用于度量卷积层输出中哪些部分对关系分类相对重要；h_{att} 是注意力池化的最终输出。

3.5 输出层

将多通道自注意力层的全局特征向量与 CNN 层的局部特征向量进行拼接，然后输入到全连接层。

为了从得到的特征中获取分类器，使用由 k 个节点组成的全连接层，k 对应关系类型的数量。然后，应用 softmax 分类器来获得每个可能的关系标签的条件概率 $p(y|x,\theta)$。最后，选取 $p(y|x,\theta)$ 中概率最大的项作为预测关系 $\hat{y}$。具体公式如下：

$$p(y|x)=\text{softmax}(W_o x+\boldsymbol{b}_o) \tag{18}$$

$$\hat{y}=\text{argmax}\ p(y|x) \tag{19}$$

其中，W_o 和 b_o 是权重参数和偏置参数，x 是上一层的输出。模型使用真实关系和预测关系的交叉熵作为损失函数，其公式如下：

$$J(\theta)=-\frac{1}{k}\sum_{i=1}^{k}y_i^{\text{true}}\log(\hat{y}_i)+\lambda\|\theta\|^2 \tag{20}$$

其中，第一项是交叉熵，又称经验风险；第二项是正则化项；y_i^{true} 表示第 i 类真实关系标签的 one-hot 编码向量表示，$\hat{y}_i$ 是通过 softmax 计算得到的第 i 类关系标签的条件概率；λ 是 L2 正则化的超参数，用于调节经验风险与正则化项之间的关系，可以更好地避免模型过拟合。模型的优化器选择 Adam。为了防止训练过程中出现过拟合现象，分别在嵌入层和 BLSTM 层加入了 Dropout。

4 实验

4.1 数据集

本文使用 CHIP2020 的评测任务二——中文医学文本实体关系抽取的数据集进行实验。该数据集由郑州大学自然语言处理实验室、北京大学计算语言学教育部重点实验室、哈尔滨工业大学(深圳)、鹏城实验室人工智能研究中心智慧医疗课题组联合构建。数据集包含儿科训练语料和百种常见疾病训练语料，其中儿科训练语料来源于 518 种儿科疾病，百种常见疾病训练语料则来源于 109 种常见疾病，包含 44 种关系类别，如表 1 所列。经过预处理后，获得约 4 万条数据，其中 32000 条用于训练，8000 条用于测试。进一步地，将训练数据按照 7.5∶2.5 的比例随机划分为训练集和验证集，其中训练集用于模型的训练，验证集则用于模型的参数调优。

表 1 中文医学关系数据集的关系类型

Table 1 Relation types of Chinese medical relational data set

关系类型	实例	关系类型	实例
预防	（麻风病，预防，利福平）	临床表现	（类癌综合征，临床表现，外周水肿）
阶段	（肿瘤，阶段，I 期）	治疗后症状	（尤因肉瘤，治疗后症状，肿瘤生长的暂时性停顿）
就诊科室	（腹主动脉瘤，就诊科室，初级医疗保健处）	侵及周围组织转移的症状	（喉癌，侵及周围组织转移的症状，颈部肿物）
辅助治疗	（皮肤鳞状细胞癌，辅助治疗，非手术破坏）	发病部位	（肿瘤，发病部位，卵巢）
化疗	（皮肤鳞状细胞癌，化疗，局部化疗）	转移部位	（肿瘤，转移部位，累及一侧或双侧卵巢）
放射治疗	（非肿瘤性疼痛，放射治疗，放射治疗外照射）	外侵部位	（侵袭性鳞状细胞癌，外侵部位，皮肤深层）
实验室检查	（HS，实验室检查，酸化甘油试验）	并发症	（登革热，并发症，横纹肌溶解症）
影像学检查	（反应性关节炎，影像学检查，X 光）	病理分型	（高苯丙氨酸血症，病理分型，苯丙氨酸羟化酶缺乏）
辅助检查	（类风湿关节炎，辅助检查，关节压痛计数）	相关（导致）	（肾间质炎症，相关（导致），特发性高钙尿症）
组织学检查	（幽门螺杆菌感染，组织学检查，组织切片法）	鉴别诊断	（阵发性室上性心动过速，鉴别诊断，都性心动过速）
内窥镜检查	（支气管哮喘，内窥镜检查，支气管镜检查）	相关（转化）	（多发性骨髓瘤，相关（转化），感染）
筛查	（急性胰腺炎，筛查，格拉斯哥预后标准）	相关（症状）	（EB 病毒感染，相关（症状），呼吸道感染）
多发群体	（SLE，多发群体，近亲发病高）	病因	（哮喘，病因，剧烈运动）
发病率	（脆性 X 综合征，发病率，2.6%）	高危因素	（HIV 感染，高危因素，成年毒品注射者）
发病年龄	（胰腺癌，发病年龄，65～75 岁）	风险评估因素	（FUO，风险评估因素，传染病接触史）
多发地区	（肺癌，多发地区，北美）	病史	（猝死，病史，不明原因的昏厥史）
发病性别倾向	（食管癌，发病性别倾向，男性）	遗传因素	（急性淋巴细胞白血病，遗传因素，同卵双胞胎）
死亡率	（成骨肉瘤，死亡率，很高）	发病机制	（HSPN，发病机制，纤维蛋白的沉积）
传播途径	（HGA，传播途径，通过蜱叮咬传播）	病理生理	（幽门痉挛，病理生理，自主神经调节功能差）
多发季节	（支原体肺炎，多发季节，秋冬季）	预后状况	（产毒素性大肠杆菌肠炎，预后状况，病程 5～10 天）
手术治疗	（皮肤鳞状细胞癌，手术治疗，传统手术切除）	预后生存率	（横纹肌肉瘤，预后生存率，80%）
药物治疗	（佝偻病，药物治疗，补充维生素 D）	同义词	（快速连续静脉肾盂造影，同义词，IVP）

4.2 实验方案

4.2.1 评价指标

本文采用精确率（Precision，P）、召回率（Recall，R）及 F1 值（F1-score）作为中文医学实体关系抽取任务的评价指标。设r_i为预设关系集合中的一个关系类型，将给定的标注结果作为真实关系标签，TP_i表示测试集中模型预测的关系类型为r_i且真实标签也为r_i的样本数量，FP_i表示预测类型为r_i但真实标签不为r_i的样本数量，FN_i表示预测类型不为r_i但真实标签为r_i的样本数量。各指标的具体计算公式如下：

$$P_i=\frac{TP_i}{TP_i+FP_i} \tag{21}$$

$$R_i=\frac{TP_i}{TP_i+FN_i} \tag{22}$$

$$F1_i=\frac{2\times P_i\times R_i}{P_i+R_i} \tag{23}$$

其中，TP_i+FP_i表示预测类型为r_i的样本数量，TP_i+FN_i表示真实类型为r_i的样本数量。

本文采用加权平均的方式来计算模型整体的精确率、召回率和 F1 值，以此作为模型整体的评价指标。加权平均的计算方式为：将各关系类型的精确率、召回率、F1 值与对应的关系标签在样本中的比例相乘，然后将所有关系类型相加。采用加权平均的方式可以解决标签不平衡的问题，但可能导致 F1 值不在精确率与召回率之间。各指标的具体计算公式如下：

$$P=\sum_{i=1}^{k}P_iW_i \tag{24}$$

$$R=\sum_{i=1}^{k}R_iW_i \tag{25}$$

$$F1=\sum_{i=1}^{k}F1_iW_i \tag{26}$$

其中，W_i为第 i 类关系类型标签在样本中所占的比例。

4.2.2 超参数设置

通过在验证集上进行参数调优实验，得到的模型超参数设置如表 2 所列。

表 2 BLSTM-MCatt-CNN 模型的超参数设置

Table 2 Hyperparameter settings of BLSTM-MCatt-CNN model

超参数名称	参数值
字向量维度	100
位置向量维度	5
LSTM 隐藏层单元数	300
自注意力通道数	30
CNN 卷积核数	200
滤波器大小	[2,5]
学习率	0.001
L2 正则化参数	0.001
Dropout 概率	0.5

4.2.3 基准模型

本文选取 6 个医学领域的实体关系抽取模型作为基准模型，具体如下。

CNN 模型[23]：利用具有最大池化的 CNN 模型进行关系抽取。

BLSTM 和 ABLSTM 模型[25]：利用双向 LSTM 模型获取语句的上下文信息，再分别结合最大池化和注意力池化进行关系抽取。

CRNN-max 和 CRNN-att[27]：分别结合 BLSTM 和 CNN 以学习文本语句的上下文信息和局部信息，并分别在最后使用最大池化和注意力池化进行关系抽取。

CBGRU[28]：结合 CNN 和 BGRU 模型学习文本语句的局部信息和上下文信息来进行关系抽取。

4.3 实验结果与分析

4.3.1 自注意力通道数的选择

多通道自注意力机制用于学习语句的深层次全局语义信

息，其通道数量的多少决定了多通道自注意力机制存储语义信息的能力。如果通道数量过少，则不能完整地捕获句子多方面的语义信息；而通道数量过多，则模型存储信息的能力过剩，提升了模型的复杂度。因此，需要通过实验确定多通道自注意力机制的通道数量，实验结果如表 3 所列。

表 3 不同自注意力通道数的实验结果

Table 3 Results with different numbers of self-attention channels

（单位：%）

r	P	R	$F1$
1	85.24	85.02	84.86
10	85.55	85.11	85.06
20	85.19	84.88	84.72
30	85.51	85.46	85.23
40	85.24	84.87	84.72

从表 3 中可以看出，当自注意力通道数从 1 增加到 10 时，模型的精确率、召回率和 F1 值都有一定的提升，说明与传统的单通道自注意力相比，增加通道数有助于模型捕获句子中多方面的语义信息，从而能较完整、深入地挖掘句子的全局语义特征，验证了多通道自注意力机制的有效性。然而，当通道数从 20 增加到 40 时，F1 值呈现先上升后下降的趋势，这表明句子的长度有限时，过多的通道数会使模型存储过多的信息，其中包含训练数据中的噪声，导致模型过拟合，从而使模型的性能有所下降。因此，通过实验将模型的自注意力通道数确定为 30，其 $F1$ 值最大为 85.23%。

4.3.2 滤波器大小的选择

在 CNN 层中，第一次池化的滤波器大小 f_1 决定了送入卷积层的信息量大小，而第二次池化的卷积核滤波器大小 f_2 可以视为匹配句子中短语的长度。两个滤波器的大小决定了 CNN 层捕获句子中局部短语特征的能力。因此，需要通过实验确定 CNN 层的两个滤波器的最佳大小，表 4 所列为两个滤波器大小在不同组合下的实验结果。

表 4 不同滤波器长度组合的实验结果

Table 4 Results of different filter length combinations

（单位：%）

$[f_1, f_2]$	P	R	$F1$
[2,3]	85.52	85.15	85.08
[2,5]	85.51	85.46	85.23
[2,7]	84.80	84.05	84.09
[3,3]	85.13	84.84	84.70
[3,5]	84.90	84.55	84.34
[3,7]	84.57	84.22	83.99

从表 4 中可以看出，如果 f_1 长度过大，可能会导致个别字符的嵌入集中在一起，使得少数字符占据了大部分区域，从而使模型忽略了其他字符的局部特征，降低了模型对局部信息的捕捉能力。而卷积核的滤波器大小 f_2 取中间值则能很好地适配句子中短语的长度。因此，模型中 CNN 层的两个滤波器的大小设定为[2,5]时最佳。

4.3.3 模型对比实验

将本文提出的 BLSTM-MCatt-CNN 模型与 6 个基准模型进行对比，结果如表 5 所列。

表 5 不同模型的实验结果

Table 5 Experimental results of different models

（单位：%）

模型	P	R	$F1$
CNN	82.03	82.21	81.68
BLSTM	82.91	82.91	82.69
ABLSTM	83.07	83.17	82.92
CRNN-max	84.19	84.24	83.94
CRNN-att	83.54	83.34	83.21
CBGRU	82.29	82.25	81.96
BLSTM-MCatt-CNN-max	**85.51**	**85.46**	**85.23**
BLSTM-MCatt-CNN -att	85.19	84.77	84.71

实验结果表明，在中文医学实体关系抽取任务上，本文提出的 BLSTM-MCatt-CNN 模型在精确率、召回率和 F1 值上比所有基准模型的效果都要好，说明本文模型分别利用多通道自注意力机制捕获全局语义特征、利用 CNN 捕获局部短语特征，并将两个特征进行融合，能有效地提升关系抽取的效果，验证了本文模型在该任务上的有效性。其中，在 CNN 层的第二次池化中采用最大池化的 BLSTM-MCatt-CNN-max 模型在精确率、召回率和 F1 值上都取得了最好的效果，$F1$ 值达到了 85.23%，超过了同样采用最大池化的当前最佳模型 CRNN-max。同时，在 CNN 层的第二次池化中采用注意力池化的 BLSTM-MCatt-CNN-att 模型的精确率、召回率和 F1 值也均已超过同样采用注意力池化的 CRNN-att 模型。上述两组模型的对比表明，在“recurrent＋CNN”网络框架的基础上引入多通道自注意力机制对句子的全局语义特征进行深入挖掘，有助于提升模型效果。而通过对比 BLSTM-MCatt-CNN-max 模型和 BLSTM-MCatt-CNN-att 模型可以发现，BLSTM-MCatt-CNN-max 模型的精确率、召回率和 F1 值都相对高一些，这与 CRNN-max 模型和 CRNN-att 模型的结果一致，表明了对于此类医学文本数据集，在 CNN 后的第二次池化使用最大池化比使用注意力池化的效果要好。

进一步分析可以看出：1）在关系抽取这个文本序列任务中，BLSTM 模型比 CNN 模型的表现更好，这是因为 BLSTM 模型能很好地捕获句子的上下文信息，从而学习句子的依赖，而 CNN 模型只能通过滑动窗口捕获句子的局部信息，无法学习句子的长依赖。2）引入了注意力机制的 ABLSTM 模型比 BLSTM 模型的关系抽取效果更好，表明引入注意力机制可以提升模型的性能。3）结合了 BLSTM 和 CNN 的 CRNN 模型相比单纯的 BLSTM 模型和 CNN 模型效果都有明显的提升，验证了“recurrent＋CNN”网络框架在医学实体关系抽取任务上的有效性，同时也证明了该结构在中文医学实体关系抽取上是有效的。4）结合了 CNN 和 BGRU 的 CBGRU 模型的性能介于 CNN 模型和 BLSTM 模型之间，表明“CNN＋RNN”框架可能不适用于关系抽取这种 NLP 任务。

结束语 本文提出了一种融合多通道自注意力机制的中

文医学实体关系抽取模型 BLSTM-MCatt-CNN。该模型在“recurrent＋CNN”网络框架的基础上结合多通道自注意力机制，将文本转换为由字向量和位置向量组成的输入特征向量后，利用 BLSTM 捕获文本句子的上下文信息和浅层语义特征，利用 CNN 捕获句子局部短语特征，再结合多通道自注意力机制捕获语句的全局语义特征，从而对医学文本的语义特征进行深入挖掘。在中文医学文本数据集上的实验对比验证了该模型的有效性。下一步工作可以考虑将医学知识等结合到模型中，以便更好地表达医学文本的语义特征。

参 考 文 献

[1] GOLSHAN P N，DASHTI H R，AZIZI S，et al. A Study of Recent Contributions on Information Extraction[J]. arXiv：1803.05667，2018.

[2] LIU Q，LI Y，DUAN H，et al. Knowledge Graph Construction Techniques[J]. Journal of Computer Research and Development，2016，53：582-600.

[3] GRISHMAN R，SUNDHEIM B. Message understanding conference-6：a brief history[C]// Proceedings of the 16th Conference on Computational Linguistics. New York：ACM Press，1996：466-471.

[4] UZUNER O，SOUTH B，SHEN S Y，et al. 2010 i2b2/VA challenge on concepts，assertions，and relations in clinical text[J]. Journal of the American Medical Informatics Association，2011，18(5)：552-556.

[5] NING S M，TENG F，LI T R. Multi-Channel Self-Attention Mechanism for Relation Extraction in Clinical Records[J]. Chinese Journal of Computers，2020，43(5)：916-929.

[6] HAN X，GAO T Y，LIN Y K，et al. More data，more relations，more context and more openness：a Review and outlook for relation extraction[J]. arXiv：2004.03186，2020.

[7] ZHAO S，GRISHMAN R. Extraction relations with integrated information using kernel methods[C]// Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2005：419-426.

[8] GUO X Y，HE T T，HU X H，et al. Chinese Named Entity Relation Extraction Based Syntactic and Semantic Features[J]. Journal of Chinese Information Processing，2014，28(6)：183-189.

[9] KAMBHATLA N. Combining lexical，syntactic，and semantic features with maximum entropy models for extracting relation[C]// Proceedings of ACL on Interactive Poster and Demonstration Sessions. Stroudsburg：ACL，2004：22-26.

[10] ZHOU J. Chinese entity relation extraction based on conditional random fields model[J]. Computer Engineering，2010，36(24)：192-194.

[11] RINK B，HARABAGIU S，ROBERTS K. Automatic extraction of relations between medical concepts in clinical texts[J]. Journal of the American Medical Informatics Association，2011，18(5)：594-600.

[12] D'SOUZA J，NG V. Ensemble-Based Medical Relation Classification[C]// 25th International Conference on Computational Linguistics. Dublin：COLING，2014：1682-1693.

[13] KIM S，LIU H，YEGANOVA L，et al. Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach[J]. Journal of Biomedical Informatics，2015，55(2)：23-30.

[14] ZENG D J，LIU K，LAI S W，et al. Relation classification via convolutional deep neural network[C]// Proceedings of the 25th International Conference on Computational Linguistics. Stroudsburg：ACL，2014：2335-2344.

[15] ZHANG D X，WANG D. Relation classification via recurrent neural network[J]. arXiv：1508.01006，2015.

[16] ZHANG S，ZHENG D Q，HU X C，et al. Bidirectional Long short-term memory networks for relation classification[C]// Proceedings of the 29th Pacific Asia Conference on Language，Information and Computation. Stroudsburg：ACL，2015：73-78.

[17] ZHU J Z，QIAO J Z，DAI X X，et al. Relation classification via target-concentrated attention CNNs[C]// International Conference on Neural Information Processing. Berlin：Springer，2017：137-146.

[18] ZHOU P，SHI W，TIAN J，et al. Attention-based bidirectional long short-term memory networks for relation classification[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2016：207-212.

[19] LEE J，SEO S，CHOI Y S. Semantic relation classification via bidirectional LSTM networks with entity-aware attention using latent entity typing[J]. arXiv：1901.08163，2019.

[20] CAI R，ZHANG X D，WANG H F. Bidirectional recurrent convolutional neural network for relation classification[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2016：756-765.

[21] ZHANG X B，CHEN F C，HUANG R Y. A combination of RNN and CNN for attention-based relation classification[J]. Procedia Computer Science，2018，131：911-917.

[22] TRAN V H，PHI V T，SHINDO H，et al. Relation Classification Using Segment-Level Attention-based CNN and Dependency-based RNN[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2019：2793-2798.

[23] SAHU S，ANAND A，ORUGANTY K，et al. Relation extraction from clinical texts using domain invariant convolutional neural network[C]// Proceedings of the 15th Workshop on Biomedical Natural Language Processing. 2016：206-215.

[24] ZHOU H W，LANG C K，LIU Z，et al. Knowledge-guided convolutional networks for chemical-disease relation extraction[J]. BMC Bioinformatics，2019，20(1)：260-273.

[25] SAHU S，ANAND A. Drug-Drug Interaction Extraction from Biomedical Texts Using Long Short-Term Memory Network[J]. Journal of Biomedical Informatics，2018，86：15-24.

[26] BAI T，WANG C，WANG Y，et al. A novel deep learning me-

thod for extracting unspecific biomedical relation[J]. Concurrency and Computation: Practice and Experience, 2020, 32: 1-11.
[27] RAJ D, SAHU S, ANAND A. Learning local and global contexts using a convolutional recurrent network model for relation classification in biomedical text[C]// Proceedings of the 21st Conference on Computational Natural Language Learning. Vancouver: CoNLL, 2017: 311-321.
[28] HE B, GUAN Y, DAI R. Convolutional Gated Recurrent Units for Medical Relation Classification[C]// 2018 IEEE International Conference on Bioinformatics and Biomedicine. New York: IEEE Press, 2019: 646-650.
[29] LIN Z, FENG M, SANTOS C N, et al. A structured self-attentive sentence embedding[J]. arXiv: 1703.03130, 2017.

ZHANG Shi-hao, born in 1996, postgraduate. His main research interests include information extraction and natural language processing.

LI Tian-rui, born in 1969, Ph.D, professor, Ph.D supervisor, is a distinguished member of China Computer Federation. His main research interests include big data intelligence, rough sets and granular computing.